Du partage des corpus à l'analyse des interactions en ligne dans des situations d'apprentissage : quelle méthodologie pour la recherche sur corpus d'apprentissage ?

Thierry Chanier, Maud Ciekanski

DU PARTAGE DES CORPUS À L'ANALYSE DES INTERACTIONS EN LIGNE DANS DES SITUATIONS D'APPRENTISSAGE : QUELLE MÉTHODOLOGIE POUR LA RECHERCHE SUR CORPUS D'APPRENTISSAGE ?

Thierry Chanier

Université Blaise Pascal

Maud Ciekanski

Université de Franche-Comté

Résumé : La recherche sur les interactions en ligne en situation d'apprentissage offre encore trop peu souvent la possibilité d'accéder aux données de recherche ayant servi à publication. Cela restreint, d'une part, la compréhension des phénomènes étudiés et, d'autre part, empêche toute réplication dans le but de comparaisons, d'analyses cumulatives ou contrastives. Cet article présente les questionnements, à la fois théoriques, techniques et méthodologiques soulevés par la conception d'un tel projet. Dans Mulce, nous défendons le point de vue méthodologique suivant : pour permettre une analyse des interactions situées, il convient de relier les différentes données issues de formations en ligne pour construire un objet d'analyse, exploitable par différentes équipes et disciplines. Le constat actuel est que les données sont souvent décontextualisées, parcellaires ou simplement inaccessibles à la communauté des chercheurs. Nous proposons de structurer les données en corpus d'apprentissage (LETEC) de façon à rendre possible son échange et la capitalisation des analyses. Le protocole de recherche, le scénario pédagogique, les interactions, productions et traces, les licences et les analyses capitalisables en sont les constituants. Le serveur Mulce permet de travailler à partir de corpus de granularités différentes que nous illustrerons à partir d'exemples issus des formations " Simuligne " et "Copéas", en indiquant les processus simples de transformation du format Mulce aux formats requis par deux logiciels d'aide à l'analyse (l'un sur les forums, l'autre sur l'alignement vidéo et transcription). Nous insistons plus particulièrement sur l'intérêt de ces outils pour l'analyse des phénomènes de polyfocalisation et d'écriture multimodale dans l'analyse des interactions multimodales. Nous conclurons sur les modes de valorisation scientifique du travail du chercheur confronté à la collecte et à la structuration de corpus d'apprentissage.

Mots-clés : Corpus d'apprentissage, corpus distinguables, interactions en ligne, interactions multimodales, réplication, partage d'outils et d'analyses.

1 Introduction

Le récent développement des environnements multimodaux d'échanges en ligne synchrones continue à susciter l'engouement de plus en plus de formateurs et d'apprenants depuis une décennie, en particulier dans le domaine de l'apprentissage des langues. Ces plateformes de formation générant des interactions complexes entre les participants renouvellent la recherche sur les interactions en ligne en situation d'apprentissage, notamment le questionnement sur les traces et leur traitement. En outre, l'intérêt pour la multimodalité (comment elle fonctionne et comment elle permet de soutenir la communication en langue-cible, voire l'apprentissage de la langue et de la culture cible) a suscité de nombreuses analyses qualitatives et quantitatives. Ces travaux ont montré que la multimodalité prend des formes différentes en fonction du type d'environnement multimodal choisi (Ciekanski et Chanier, 2008) : elle peut être uniquement *verbale* (les environnements *audio-synchrones* intégrant audio et clavardage, Jepson, 2005), ou combiner une communication *verbale* et *non-verbale* comme dans les environnements de vidéoconférence intégrant la vidéo, l'audio, le clavardage et l'iconique (Wang, 2004), et comme dans les environnements *audio-graphiques synchrones* intégrant en plus de ces modalités texte, image et graphique (Chanier et Vetter, 2006 ; Lamy, 2007). Ces analyses reposent souvent sur la collecte de données issues de la réplication d'expériences en formation concernant la plupart du temps un seul type de dispositif multimodal.

Par ailleurs, la méthodologie de recherche utilisée habituellement dans l'analyse des interactions en ligne synchrones et asynchrones ne prend en compte qu'une seule partie des données, voire ne porte que sur un seul type d'interaction (quelques forums, quelques extraits de séances synchrones), et ne les organise pas de façon structurée. Ainsi l'absence de pratiques et de guides concernant la constitution des corpus et l'absence de constitution de corpus organisés à des fins d'évaluation des approches et des stratégies d'apprentissage freinent la recherche dans le domaine, d'autant plus qu'il s'agit de corpus qui peuvent être perçus comme coûteux à collecter et à annoter, notamment pour les corpus multimodaux, en raison des multiples niveaux d'annotations requis (les diverses propriétés de la parole, les diverses propriétés des actions réalisées dans les collecticiels, la location des acteurs, etc.) selon les différents niveaux de signification visés (schèmes communicatifs, schèmes collaboratifs, etc.).

Parmi les communautés de chercheurs travaillant sur les interactions en ligne, notamment celle sur l'apprentissage collaboratif (CSCL, *Computer-supported Collaborative Learning*), il existe un intérêt grandissant pour les questions relatives à la méthodologie de structuration de corpus et sur la définition de formalismes pivots en vue des échanges et du partage des données, des analyses et des outils. Pourtant, cet intérêt est encore insuffisamment partagé dans le domaine de l'apprentissage des langues en ligne, malgré un intérêt grandissant pour la communication multimodale, dans des environnements variés qui nécessiteraient des possibles comparaisons pour mieux appréhender la nature de la multimodalité et ses potentialités en termes d'apprentissage.

Dans le projet Mulce (*MULtimodal Corpus Exchange*), nous définissons la notion de corpus d'apprentissage en identifiant l'information qu'il doit contenir, structurée de façon à rendre possible son échange et la capitalisation des analyses. Le protocole de recherche, le scénario pédagogique, les interactions, productions et traces, les licences et les analyses capitalisables en sont les constituants. La structuration des corpus d'apprentissage offre donc aux chercheurs de nouvelles perspectives. En effet, le format adopté pour encoder un corpus d'apprentissage peut être automatiquement traduit dans d'autres formats requis pour des outils

d'analyse développés par d'autres communautés travaillant sur les interactions et le langage (EIAH, CSCL, TAL).

Nous illustrerons nos propos en définissant dans un premier ce que nous entendons par la recherche sur corpus pour l'analyse des interactions en ligne en situation d'apprentissage. De cette définition découle notre approche méthodologique telle que mise en œuvre dans le projet Mulce. A partir d'exemples issus d'interactions verbales et non-verbales dans deux formations (l'une asynchrone, Simuligne, l'autre synchrone, Copéas), nous aborderons la notion de granularité des corpus en vue de leur partage et de leur analyse et nous indiquerons le processus de transformation du format Mulce aux formats requis par deux logiciels d'aide à l'analyse (l'un sur les forums, l'autre sur l'alignement vidéo et transcription), issus des recherches sur l'analyse des traces en EIAH, pour de possibles analyses, elles-mêmes partageables. Nous aborderons enfin les conditions permettant l'échange de tels corpus et leur accès libre.

2 De l'intérêt du corpus dans les recherches du domaine EPAL

2.1 Interprétations multiples sur données incomplètes

Revenons sur le cas, trop rare, de données issues d'une seule situation de formation en ligne ayant donné lieu à une série d'interprétations contrastées par des auteurs différents. La publication de citations récentes, abondant dans le sens de certaines interprétations, ainsi que l'étalement dans le temps des différentes publications permettent de soulever certaines considérations épistémologiques, autour d'un cas jugé important par la communauté de chercheurs, relié à un seul ensemble de données.

En 1997, une formation en ligne met en rapport une classe de lycée français et une d'étudiants aux États-Unis. Lors des communications exolingues par courriel autour de tâches interculturelles survient un incident critique entre les deux communautés. Ces phénomènes, objets d'une grande attention dans les recherches sur l'interculturel (Audras et Chanier, 2008), est d'abord expliqué par l'auteur de l'expérimentation comme étant le résultat d'incompréhensions de nature linguistique (Kern, 2000). Une description sommaire de la situation de formation et un ensemble fragmentaire des données d'interactions sont transmises par l'auteur à deux autres chercheurs, Kramsch et Thorne, qui n'ont pas participé au projet initial. Ceux-ci donnent une interprétation différente (Kramsch et Thorne, 2001) en invoquant la prise en considération de compétences de communication spécifiques à l'Internet, qui doivent être étudiées en tant que telles à l'échelle de la globalisation des communications. Thorne (2003) poursuit cette réinterprétation en l'orientant vers les cultures et les littératies des participants (voir aussi (Kern, Ware et Warshauer, 2004) pour cette mise en perspective des différentes étapes). Basharina (2007 : 84), reprend une partie de cet historique des publications et appuie la dernière interprétation :

> *Thorne (2003) argues that online and other activities […] represent the "culture-of-use" of an artifact (p. 40). He found that the activity of online interaction was different for the French than it was for the Americans, in part because the Internet communication was used differently in each case; e.g., French students were communicating through a surrogate (the teacher who was sending their messages). Thorne concludes that radically different cultures-of-use of Internet communication was one of the major reasons for the tension between the French and American students.*

Qu'entend-on exactement par "culture d'usage de l'Internet" ? Veut-on désigner des formes de communication en ligne qui seraient différentes entre adolescents étasuniens et français ? Les jeunes français dans le cadre de leurs loisirs utilisaient-ils en 1997 l'Internet, par exemple les messageries instantanées et le courriel, d'une manière fondamentalement différente de

celles des étasuniens ? Ne pourrait-on pas plutôt parler de cultures institutionnelles différentes ? Car la classe de langues au secondaire en France, avec ses programmes, sa culture de formation des enseignants, ses protocoles d'échanges entre enseignant et élèves était et reste bien différente de celle de la culture d'apprentissage en langues à l'université aux Etats-Unis. Quant à l'Internet, son usage à l'époque dans les institutions françaises du secondaire était fortement bridé, contrôlé, ce qui pourrait expliquer le choix contre-productif de l'enseignant de faire passer tous les messages des lycéens par son courriel individuel. Que penser en outre de la compétence interculturelle de l'enseignant ? Était-il sensibilisé à l'occurrence et la gestion des incidents critiques, compétence dont l'importance a été soulignée indépendamment de celle concernant la communication sur Internet ?

En l'absence de données disponibles sur le scénario pédagogique convenu entre les deux enseignants, de questionnaires remplis par les apprenants, voire les enseignants, sur leurs usages de l'Internet, leurs pré-conceptions sur l'interculturel, sur le déroulement de la formation (observations, ensemble des échanges en ligne, voire discussions en présentiel), beaucoup d'interprétations divergentes sont possibles.

Notre propos n'est nullement de critiquer le protocole expérimental mis en place par Kern dans une formation exploratoire et originale en 1997. C'est plutôt l'exégèse des articles publiés sur ce sujet, et d'autres du même domaine de recherche se citant de façon croisée, qui oriente le lecteur vers une interprétation stéréotypique, que l'on pourrait trouver aussi culturellement marquée (vision nord-américaine des différences entre pays de communication sur l'Internet). L'élaboration d'une recherche sur des données bien trop parcellaires tend à lui conférer des caractères impressionnistes (*impressionistic* en anglais) où tout un discours scientifique s'élabore à partir d'exemples en partie décontextualisés. On a l'impression de retrouver, sur un plan épistémique, l'opposition fondamentale entre sciences de *l'exemplum* et sciences du *datum*, telle que rappelé par Laks (2008) en linguistique et particulièrement en phonologie où il défend le travail systématique à partir de corpus.

2.2 Le paradigme corpus

Dans notre article (Reffay et al., 2008), nous avions introduit la notion de corpus telle que perçue dans différents domaines des sciences du langage précurseurs en la matière, à savoir le traitement automatique des langues, la linguistique textuelle et les interactions orales. Avant de nous concentrer sur le domaine des interactions en ligne en situation d'apprentissage, parcourons un domaine connexe en pleine expansion, à savoir les corpus en apprentissage / enseignement des langues.

Le terme "corpus" étant utilisé dans des sens très différents, comme nous le verrons, il est important de rappeler nos choix. Dans l'article précité, une définition de Bommier-Pincemin nous avait servi de point de départ. Nous aurions pu tout aussi bien partir de celle, convergente, de Rastier (2005 : 32) donnée à l'occasion d'une conférence sur "la linguistique de corpus". Nous préférons aujourd'hui définir cette notion sous forme de paradigme, au sens anglo-saxon, à savoir de modèle explicatif de l'objet considéré. Le paradigme corpus comporte les quatre points indissociables suivants.

- **Recueil systématique des documents** liés à l'objet d'étude. La couverture et la taille sont alors des indicateurs de la systématicité du recueil. Les documents ne se limitent bien sûr pas à des textes, sauf si on comprend ce mot dans un sens élargi comme le font Halliday (1989) ou Baldry et Thibault (2006), à savoir des documents multimodaux (sur le caractère multimodal des interactions en ligne, voir (Betbeder et al., 2008)).
- **Description du contexte**. Le contexte couvre, bien entendu, les métadonnées décrivant en termes précis les caractéristiques de l'œuvre, ses acteurs (collecteurs, contributeurs, etc.) telles que recommandées, par exemple, dans le

standard OLAC (2008) de l'*Open Language Archives Community*. Bien au-delà, le contexte est d'abord celui de l'expérimentation d'apprentissage, des interactions (au sens de Goodwin et Duranti (1992), cité dans (Reffay et al., 2008)), de la recherche, cause de cette expérimentation. Il est souhaitable que ce contexte soit décrit suffisamment précisément de façon à pouvoir être mis en relation avec le recueil de documents précités et offrir des paramètres d'analyse. En comparaison avec la linguistique de corpus s'appuyant sur des textes, le terme "contexte" est à rapprocher des notions de "cotexte" et "intertexte".

- **Organisation et instrumentalisation en vue de traitements**. Un corpus s'élabore en vue d'analyses multiples. Même si certaines comportent des phases manuelles, elles sont toujours assistées par des outils, à défaut d'être entièrement automatiques. Les travaux de la communauté CATCOD (2008) sur les corpus oraux montrent bien qu'une équipe de recherche aura fréquemment recours à plusieurs outils dans un même projet. Se pose alors la question de la transmission des données produites d'un outil à l'autre. Pour y répondre, le projet européen SACODEYL (2008), producteur de corpus oraux à des fins pédagogiques sur le parler des adolescents de différents pays, a même constitué une chaîne de traitement complète dont les outils peuvent être déployés du laboratoire à la salle de classe. Par ailleurs, tout le monde s'accorde sur le fait qu'un corpus s'organise en vue de permettre aussi bien des analyses qualitatives que quantitatives, comme le soulignent O'Keefe et al. (2007 : 2) dans leur ouvrage sur les corpus de langues à destination pédagogique. De ces considérations découle le fait que tous les documents recueillis et organisés doivent être numérisés dans des formats ouverts, adaptables à différents outils, que les données provenant de ces documents ou du contexte d'origines textuelle ou transcrites sous forme textuelle, doivent être organisées dans des langages de balisage, ouverts aux traitement, comme XML, et doivent être structurées suivant des schémas standard comme la TEI, ou, à défaut, des schémas accessibles à tous (voir (Reffay et al., ibid) pour les références techniques).
- **Dispositions en vue de l'échange et du partage**. Pour qu'une démarche scientifique puisse se dérouler avec ses phases d'analyses multiples, réanalyses, discussions contradictoires, il est nécessaire que le corpus soit organisé en vue de l'échange et du partage au sein des communautés de chercheurs et d'enseignants. Le corpus, ou plutôt la banque de corpus, sera en accès libre sur un serveur indexable par les autres serveurs de la Toile suivant des protocoles standard, comme celui des archives ouvertes (Chanier, 2004), repris par OLAC (ibid). Sur ce site seront clairement indiqués, le cadre d'utilisation du corpus sous forme de licence, tout comme les questions afférentes à l'organisation de l'expérimentation / recherche où ont été collectées les données dans le respect de l'éthique (Oates, 2006), leur collecte, pré-traitement (anonymisation), dépôt dans le respect des droits (Baude et al., 2005 : chap 2). Le type d'accès au corpus n'est que la première caractéristique des notions de partage et d'échange. Les autres concernent les fonctionnalités de dépôt sur le site de nouveaux corpus, de nouvelles analyses en rapport avec les corpus existants, de référencements et d'accès aux outils d'analyse associés.

Par sa définition du corpus, Rastier (ibid) écartait de son objet d'étude les corpus de mots, d'attestations ou d'exemples et les corpus de fragments. De même, le paradigme corpus décliné ici permet de différencier corpus et base de données, et de ne pas considérer comme

constituants d'un corpus les recueils de documents authentiques numérisés, au contraire de la position prises par plusieurs chercheurs dans le colloque "Des documents authentiques aux corpus oraux : questions d'apprentissage en didactique des langues" (CRAPEL, 2007).

On pourrait croire que le respect des différents points du paradigme corpus engendre un niveau tel de complexité et de lourdeur de tâches de nature à décourager toute équipe de recherche de constituer une telle œuvre. Paradoxalement, dans le domaine des apprentissages ou de la didactique, ce sont les chercheurs qui ne s'inscrivent pas dans ce paradigme, ni dans la méthodologie afférente qui doutent fortement de la possibilité de constituer des corpus :

> *Les programmes de formation linguistique destinés aux étudiants allophones intégrant l'université française s'intéressent […] à la compréhension orale des discours enseignants. Ils tentent pour cela de s'appuyer sur l'enregistrement de cours magistraux ou de travaux dirigés. Mais la transformation de ces enregistrements en supports pédagogiques est loin d'être aisée. Ce type de discours résiste aux habitudes établies par la didactique du FLE en matière d'enseignement de l'oral, […], leur utilisation dans la classe pose encore beaucoup de questions. Ce qui explique la quasi inexistence des discours académiques dans la panoplie des documents authentiques en FLE. (Parpette, 2007)*

Pourtant, d'autres chercheurs de la même communauté disent (Detey et al., 2007) que l' "*on doit s'interroger sur la possible utilisation des corpus oraux existants, généralement conçus sur la base de motivations initialement linguistiques, au sens large du terme*".

Si maintenant, on regarde du côté des chercheurs s'inscrivant dans le paradigme évoqué, outre le projet SACODEYL déjà cité, on trouvera en particulier la banque de corpus MICASE (2009) qui a précisément réussi sur l'anglais académique, ce qui était jugé inatteignable pour une partie du milieu FLE. De plus Pérez-Llantada (2009) a monté une expérimentation dans laquelle, à partir de MICASE, elle mesure l'impact de l'utilisation de tels corpus sur l'apprentissage de diverses compétences en compréhension en L2 et, en production, lorsque les apprenants les utilisent dans des scénarios dédiés. Ces travaux s'inscrivent dans la, déjà longue, tradition du milieu TALC (*Teaching and Language Corpora*) (O'Keefe et al., ibid).

S'inscrire dans ce paradigme permet au milieu des chercheurs et praticiens opérant sur des situations d'apprentissage ou d'enseignement, soit d'exploiter les banques de corpus à des fins de recherche sur l'apprentissage / acquisition des langues, soit d'enseigner ou d'aider à apprendre des langues. De plus, corpus, méthodologie et outils servent de cadre à la formation de jeunes chercheurs qui demain étendront ces banques de corpus.

L'enjeu est donc ici de faire entrer dans ce paradigme le "genre" des interactions en situations d'apprentissage en ligne, avec des études portant sur différents types de "discours" (pour reprendre la terminologie de Rastier (ibid : 34)). Car, même si nos exemples renvoient majoritairement à des situations d'apprentissage des langues, notre propos vise bien tout domaine d'apprentissage impliquant des interactions en ligne, que cela soit en géographie ou en mathématique pour ne prendre que deux exemples très différenciés. Pour ce faire, nous introduisons un nouveau type de corpus, dénommé *corpus d'apprentissage* (LEarning & TEaching Corpus) ou LETEC, de façon plus concise. Notre propos concerne, non seulement les sciences du langage, mais également le domaine des environnements informatiques pour l'apprentissage humain (EIAH) et de l'apprentissage collaboratif (CSCL).

3 Vue d'ensemble sur le concept de corpus d'apprentissage

Dans cette section nous brossons une vue d'ensemble d'un corpus d'apprentissage. Pour une information détaillée sur le sujet, particulièrement les modèles de structuration choisis, le lecteur pourra se reporter à (Reffay et al, 2008), (Reffay et Betbeder, à paraître). Donnons une première définition

Un corpus d'apprentissage (LETEC) assemble de façon systématique et structurée un ensemble de données, particulièrement d'interactions, et de traces issues d'une expérimentation de formation partiellement ou totalement en ligne, enrichies par des informations techniques, humaines, pédagogiques et scientifiques permettant leur analyse en contexte.

3.1 Constituants du corpus

La figure 1 de gauche schématise les différentes parties du corpus.

- Le dispositif pédagogique peut-être librement décrit, mais il est préférable de le faire de façon détaillée en précisant le scénario pédagogique, les différents rôles des participants, en particulier des apprenants et enseignants, ainsi que les environnements technologiques retenus avec leurs fonctionnalités, et leurs caractéristiques dédiées aux interactions.
- De la même façon, si l'expérimentation inclue un protocole de recherche, le rôle des chercheurs, le séquencement des activités afférentes (administration de questionnaires, entretiens, etc.) seront utilement décrits.
- Les deux parties précédentes correspondent à ce qui était prévu avant le déroulement de la formation, à un modèle donc. Suivant la terminologie des langages objets, le modèle s'instancie lors de l'acte pédagogique (avec tous les changements inopinés afférents). La partie instanciation assemble donc, d'une part, les enregistrements des interactions des participants (sous formes textuelle, audio ou vidéo), leur productions individuelles (tels les travaux écrits ou oraux, les journaux de bord) et, le cas échéant, les traces système (temps de connexion, statistiques de participation, etc.). Elle regroupe d'autre part, le cas échéant, les questionnaires remplis, les enregistrements d'entretiens, les matériaux afférents (grille d'entretien, matériaux pour auto-confrontation, etc.).
- La partie publique de la licence donne accès aux licences d'utilisation du corpus par la communauté de chercheurs et praticiens (Mulce a choisi une licence Creative Commons) et les formulaires de contrat d'éthique remis aux participants. La partie privée de la licence, n'est pas directement intégrée au corpus, mais conservée par le responsable du corpus. Elle incorpore notamment les patronymes et coordonnées des participants, ainsi que les contrats signés.
- Les analyses ne font pas en général pas partie du corpus d'apprentissage, mais seront adjointes ultérieurement sous forme de corpus distinguables (voir ci-après). Quant au cas intermédiaire des transcriptions des enregistrements vidéo ou audio, nous les avons intégrées dans les corpus déposés dans le serveur Mulce, en sachant qu'elles peuvent être recommencées ou modifiées. Le schéma de la figure 1 (gauche) fait apparaître la partie analyse comme l'objectif orientant l'ensemble de l'effort de collecte et d'organisation.

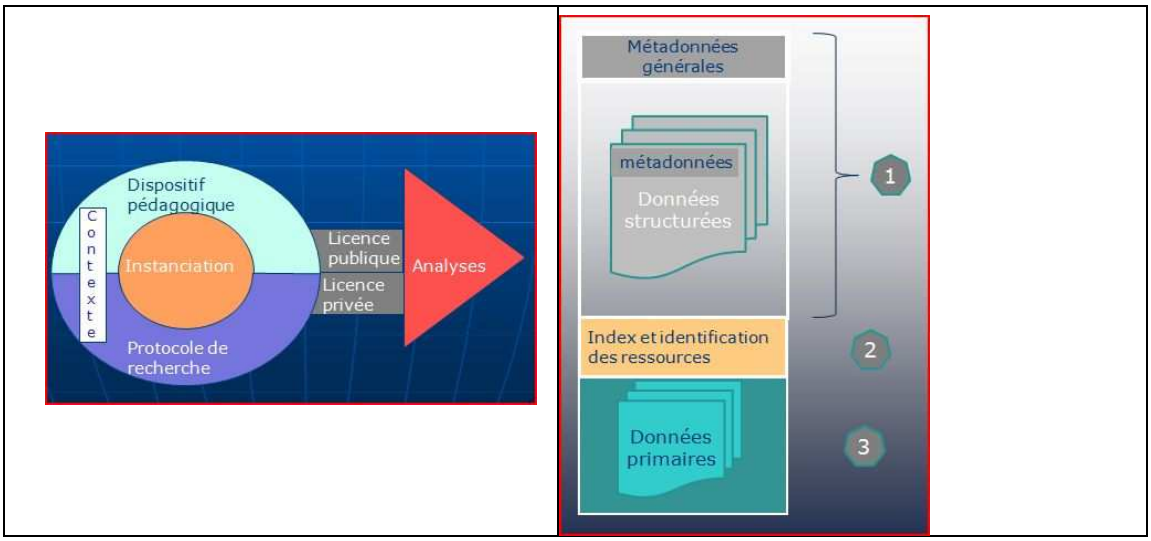

Figure 1 - Les grandes parties d'un corpus d'apprentissage LETEC (figure de gauche) et sa structuration (droite)

3.2 Structuration du corpus

La partie droite de la figure 1 indique comment toutes les données et informations du LETEC s'organisent. En 3, figurent les données primaires des différentes parties (consignes pédagogiques, traces des interactions, questionnaires et entretiens, etc.), dans les formats nécessaires à leur conservation et utilisation. Le qualificatif "primaire" adjoint à ces données, suivant la terminologie courante des corpus en linguistique, est quelque peu abusif, puisqu'une partie des formats des documents, (textes, vidéogrammes, audiogrammes) pourront être transformés et les documents eux-mêmes anonymisés, plus ou moins profondément, suivant les exigences précitées.

En 2, sont regroupées les index, identifiants et informations résumées sur chacune des ressources de la partie 3. Cette partie est structurée, laissant ainsi apparaître les groupes de ressources correspondant, par exemple, à un ensemble de consignes pédagogiques, un ensemble de données pour des entretiens semi-directifs, ou un ensemble de fichiers permettant d'aligner transcriptions et vidéo (cf. exemple plus loin).

La partie 1 est entièrement structurée, avec le langage de balisage XML, suivant un ensemble de schémas. Elle contient un premier ensemble de métadonnées générales du corpus, au format OLAC, qui lui permette d'être identifié, depuis le serveur Mulce, par les internautes et les serveurs moissonneurs de la Toile, suivant les protocoles déjà mentionnés. Figurent ensuite, pour chaque partie du corpus, les informations structurées afférentes. Pour ce qui concerne les interactions, nous avons mis au point une structure, correspondant au schéma *Mulce-struct* (Reffay et al., 2008 : s.3.3.2), dans laquelle sont encodés de façon homogène, les messages de courriels, de forums, de blogues les modalités des environnements audio-graphiques (audio, clavardage, iconique, actes de production dans des tableaux blancs, traitement de texte partagé, carte conceptuelle, blogue, etc.). Le contenu des actes, résultats de transcription (Chanier et Vetter, 2006) ou des traces système, est en correspondance avec les environnements technologiques et les participants. Une autre partie de la structure code, justement les informations ethnographiques et éléments de bibliographie langagière, essentiels pour les analyses linguistiques ultérieures (Belz et Vyatkina, 2008 : 45). De même, si les collecteurs et éditeurs du corpus ont choisi de structurer les informations concernant le dispositif pédagogique, alors chaque acte attaché à une interaction peut-être automatiquement mis en rapport avec le contexte pédagogique (activité, rôle des participants, consignes, etc.). Pour les premiers corpus d'apprentissage déposés dans le serveur Mulce (cf.

tableau 1), nous avons choisi d'utiliser les modèles standard développés par la communauté *IMS Global Learning Consortium* (IMS, 2009) modèles par ailleurs souvent utilisés dans les cursus de formation d'ingénieur pédagogique dans les phases de conception de situations d'apprentissage.

Signalons que ces trois parties sont enveloppées dans un container (*content packaging*) correspondant à l'un des format prescrit par IMS, la partie 1 étant dénommé le *manifeste*. C'est ce même format qui est souvent utilisé pour échanger des ressources pédagogiques entre plates-formes de téléformation.

À la lecture de cette partie offrant une vue d'ensemble sur les corpus d'apprentissage, le lecteur aura sans doute pu aisément faire le rapprochement avec les différents points du paradigme corpus présenté ci-avant. Revenons sur le point sur l'organisation et de l'instrumentalisation en vue des traitements. Le premier intérêt de ces efforts de structuration est de pouvoir offrir au chercheur accédant au serveur de la banque de corpus Mulce un environnement de travail lui permettant d'effectuer des fouilles *intra* ou *inter corpus*, comme c'est le cas dans le serveur Clapi (2009), banque de corpus dans le champ des interactions verbales. Le second intérêt est de permettre de réutiliser aisément ces données structurées pour mener des analyses avec des outils développés par la communauté de recherche sur les interactions en ligne, comme nous allons le développer ci-après. Reste la question du coût d'organisation et de structuration de tels corpus. Elle sera évoquée en fin d'article.

	Simuligne	Copéas
Objectif Pédagogiques	FLE en formation continue pour anglais	Anglais pour Master2 FOAD
Institutions	UFC, OU	OU, UFC
Participants	- 1 coordinateur - 10 natifs (UFC), - 40 apprenants (OU) - 4 tuteurs (OU). - 4 groupes de 12 - 1 groupe global (60)	- 14 apprenants (UFC) - 2 tuteurs (OU) - 2 groupes de 7+1
Environnements technologiques	Asynchrone (WebCT)	Synchrone:(Lyceum) Asynchrone:(WebCT)
Interactions Devoirs rendus	- 2686 mess. forum, - 4062 courriels - 5680 tours de clavardage - 93 doc. textuels, - une image - 28 fichiers audio	- 5506 tours de parole audio (8h29 en temps cumulé) - 1529 tours de clavardage - 16 séances Lyceum
Productions affichées	342 pages web incluant 115 images et 44 fichiers audio	Documents, cartes conceptuelles et tableaux blancs
Ressources pédagogiques	guide apprenant guide tuteur guide natifs	guide apprenant guide tuteur
Scénario	28 activités réparties en 7 étapes / 12 semaines,	8 activités sur 10 semaines
Questionnaires, Entretiens	12 questionnaires apprenants,	- 14 quest. app., - 7 entretiens, - 9 *Critical Event Recall* (8 app., 1 tuteur)

Taille	Total : 650 Mo : - 30 000 fichiers répartis dans 2708 dossiers	Total : 35,3 Go : - 37 vidéos (27h) - 512 autres fichiers dans 117 dossiers. - 180 000 lignes de traces et transcription dans Mulce-struct
Cession droits, contrat éthique	Cession droit (oui), contrat éthique (non)	Oui

Tableau 1 • Description synthétique des ensembles de données de 2 corpus d'apprentissage déposes dans le serveur Mulce. UFC (Univ. de Franche-Comté), OU (Open Univ.)

4 Analyses et corpus distinguables

Nous avons parlé d'analyses pouvant s'opérer à partir de fouilles sur la totalité du corpus d'apprentissage à partir du serveur Mulce, permettant, par exemple, de calculer la durée moyenne des actes de parole par acteur dans l'environnement audio-graphique utilisé dans la formation Copéas (Vetter et Chanier, 2006). Dans cette section, nous discuterons d'analyses pouvant être assistées à l'aide d'outils développés par des chercheurs du domaine des interactions en ligne. Pour ce faire, il nous faut au préalable introduire un type de corpus de granularité inférieure à celle du corpus d'apprentissage.

4.1 Des corpus de granularités différentes

Comme le montre les exemples du tableau 1, un corpus d'apprentissage correspondant à une expérience de formation est un méga corpus comportant une trop grande quantité de données pour pouvoir offrir des objets aisément analysables. En outre, ces données sont de nature hétérogène et relèvent de phénomènes complexes et dynamiques. Il devient alors nécessaire de travailler à partir d'unité intermédiaire, d'où la constitution de *corpus distinguables* (Reffay et al., 2008 : s 2.6). En partant d'un corpus d'apprentissage, que l'on qualifiera de *corpus global*, il est possible de produire des corpus distinguables, chacun correspondant au grain habituellement retenu par un chercheur pour y accomplir une analyse sur un phénomène précis (cf. exemples dans la section suivante). Le corpus distinguable est tout à la fois un sous-corpus du corpus d'apprentissage et un corpus en soi. Son container est de même format que celui d'un corpus global (cf. figure 1, partie de droite). Au contraire du dernier, il est facilement téléchargeable sur un ordinateur personnel. Le chercheur dispose alors d'un ensemble comportant une description structurée du corpus, contextualisé par rapport au corpus global (sous forme de commentaires libres et d'index précis renvoyant sur chacune des sous-parties d'un corpus global), des outils d'analyse associés et un ensemble de données prêtes à l'analyse ou contenant déjà des résultats d'analyse. Enfin des liens relient un corpus distinguable à son corpus global et, le cas échant, à d'autres corpus distinguables pour des analyses inter-corpus.

Les corpus distinguables constitués dans Mulce répondent à trois objectifs variés, que nous distinguerons en trois types de corpus distinguables :

- associer publication scientifique et données (type 1) ;
- rassembler des données prêtes à l'analyse avec mise en forme pour outils/logiciels libres (type 2) ;
- partager des analyses avec des outils associés (type 3).

Nous illustrons notre propos à l'aide d'exemples issus de nos formations.

4.2 Associer publication scientifique et données (type 1)

Lorsque, dans les différentes revues ou colloques organisées par les disciplines travaillant dans le domaine des interactions en ligne, un auteur soumet un article mettant en mots les résultats de sa recherche construites à partir de données, le comité scientifique n'a aucun moyen d'accéder à ses données, ni donc de vérifier la qualité du traitement de l'auteur. De même les lecteurs de l'article publié n'ont aucun moyen systématique d'obtenir toutes les informations ayant suscité ladite publication. Ils ne peuvent ni refaire des analyses sur ces données, ni répliquer l'expérience.

Conscient de ses limitations, nous avions dès 2005, commencé à associer à nos dépôts de publications dans les archives ouvertes, des fichiers de données (le cas échéant). Aujourd'hui grâce à la notion de corpus distinguable, nous avons pu reprendre les publications associées aux deux corpus globaux Simuligne et Copéas (cf. tableau 1) et construire les corpus en rapport. Ainsi, à partir de l'article (Reffay et Chanier, 2003), qui modélisait la structure des groupes d'apprentissage en ligne dans la formation Simuligne au travers de leurs interactions, nous avons rassemblé et structuré l'ensemble des informations, répondant ainsi aux questions souvent adressés par des lecteurs sur nos calculs, données et outils provenant de la communauté des réseaux sociaux (voir (Reffay, 2009) pour lire le manifeste du corpus distinguable).

Autre exemple, qui montre comment un corpus distinguable peut offrir des perspectives supplémentaires à celle d'un article. Notre collègue, T. Lewis, qui avait endossé le rôle de tuteur, enseignant d'anglais de spécialité, lors de la formation Copéas, avait publié un article (Lewis, 2006) de nature auto-réflexive sur ce rôle. Nous avons, avec lui, repris l'ensemble des données à partir duquel il avait construit son analyse et les avons complétées de données supplémentaires, toutes extraites du corpus global, indiquant le point de vue des apprenants. L'explicitation de l'ensemble dans le corpus distinguable (dont le manifeste est (Lewis, 2009)) montre l'appréciation différente des apprenants sur la qualité des processus collaboratifs. Ce type de support (le corpus distinguable), associé à un processus de confrontation croisée avec des tiers est peut-être un moyen de pratiquer plus en profondeur la réflexivité, comme cela se fait déjà en psychologie du travail, que cela soit pour les apprenants ou les enseignants (Chanier et Cartier, 2006). On notera que le corpus distinguable vient conforter ici une *recherche de type purement qualitative*, ce qui n'exclut donc pas une confrontation avec les données.

Si l'association de publications d'articles et de données n'en est encore qu'à ses prémisses dans notre domaine, elle devient systématique dans un nombre croissant de champs disciplinaires. Elle est obligatoire depuis des années en médecine expérimentale, où les articles sont reliés par un hypertexte aux données, en biologie pour le décryptage du génome humain, où le préalable à la soumission en vue de publication est le dépôt de la nouvelle séquence dans des banques de données mutualisées. Elle connaît un essor récent dans les sciences sociales qui développe un nouveau paradigme de publication des travaux scientifiques, autour de la notion d'ensemble de données pour la réplication :

> ***Replication data sets*** *include the original data and any other information needed to reproduce the numerical results in a published work. […] making publicly available a replication data set for each of their empirical articles or books. Citation credit should be apportioned both for the original article and separately for the data.* (Gary, 2007 : 145)

Afin de mettre ces propos en application, la communauté de chercheurs afférent à développer le réseau Dataverse (2009) qui relie les archives de dépôts de données de recherche et offre des outils pour ces serveurs, ainsi que pour les revues désirant changer leur processus de soumission. Ce milieu scientifique offre ainsi une réponse effective à un point fondamental pour tous les chercheurs travaillant sur des corpus, à savoir la reconnaissance scientifique et la valorisation des carrières des chercheurs opérant en sciences du *datum*.

4.3 Rassembler des données prêtes à l'analyse avec mise en forme pour outils/logiciels libres (type 2)

Un corpus d'apprentissage est un objet complexe à partir duquel il est possible d'extraire les matériaux pour mener de nombreuses analyses. Les collecteurs et éditeurs du corpus n'en ont souvent effectué qu'un petit nombre au moment du dépôt. Comme nous le verrons dans la section suivante, ils ont alors les moyens d'en effectuer de nouvelles. Si celles-ci ne sont accomplies que par les créateurs du corpus, elles laisseront en latence de nombreuses possibilités, qui pourraient être croisées avec des données provenant d'autres formations. C'est la raison pour laquelle nous avons imaginé une situation intermédiaire et l'avons mis en œuvre en construisant des corpus distinguables de type 2. Pour ce faire, il convient d'identifier des objets dignes d'intérêts pour les chercheurs du domaine, en extraire les données, les documenter et contextualiser par rapport au corpus global, transformer les formats de ces données pour les mettre aux formats des logiciels libres développés par la communauté.

Le premier type d'objet considéré concerne les forums de discussion. Cette modalité de communication est au cœur de nombreuses études en sciences du langage, EIAH ou CSCL. Des communautés se sont naturellement constituées comme celle de Calico (2009), où enseignants des écoles et chercheurs en EIAH se coordonnent pour, tout à la fois, spécifier les fonctionnalités des outils d'analyse, les développer et les rendre disponibles à la communauté sur des serveurs qui jouent également la fonction de banque de forums (Bruillard, 2008). Une partie de ces outils permettent de visualiser et quantifier contributeurs aux forums, fils de discussion, organisation temporelle des échanges. D'autres permettent de créer des mini-lexiques décrivant des thématiques et d'observer la reprise de ces thématiques dans les messages et les fils.

Partant du corpus global de Simuligne, nous avons donc extrait l'ensemble des forums, issus de la simulation globale en langue et de l'activité Interculture. Les 100 forums correspondants ont été extraits du corpus distinguable, transformés par des traitements automatiques du format Mulce-struct au format XML-forum de Calico, regroupés dans un corpus distinguable (Chanier, 2009), documentés avec les scénarios pédagogiques et les consignes d'utilisation des outils Calico. Ces forums sont tout à la fois disponibles sur le site de Calico (ibid), donc librement accessibles et prêts pour les outils de traitement du site et dans le corpus distinguable, prêts pour une analyse inter-forum avec, par exemple, des concordanceurs ou autre outils de traitement du langage.

Dans l'autre corpus d'apprentissage correspondant à la formation Copéas, l'objet d'étude se focalise sur les interactions multimodales en ligne. Une étape initiale, souvent indispensable, pour mener des analyses, consiste à pouvoir aligner transcription et vidéo dans des outils dédiés, en l'occurrence ici Tatiana (2008). Nous avons donc extrait du corpus les transcriptions des vidéogrammes, les avons transformées par traitement automatique du format Mulce-struct vers celui de Tatiana et constitué de la même façon que précédemment un corpus par vidéogramme. La section suivante donne plus de détails sur le sujet et illustre le fait que partant d'un corpus distinguable représentant une session ou une fraction de session d'apprentissage en ligne, on peut dériver des corpus dans lesquels sont déposés des analyses.

4.4 Partager des analyses avec des outils associés (type 3)

Afin de pouvoir analyser les interactions et tenter de comprendre les phénomènes qui se sont déroulés durant une formation en ligne, il est indispensable de travailler sur « l'organisation, la modélisation et la conceptualisation des traces d'activité, de leur représentation et de leur traitement » (Settouti et al, 2006). La difficulté rencontrée reste l'obtention d'une représentation optimale des données rendant observable l'interaction humaine et en permettant l'analyse. En effet, étant donné le volume important de données

traitées (cf. tableau 1) et la complexité des interactions analysées, les requêtes présentées « à plat », par exemple par une base de données, mettent peu en évidence les phénomènes intéressants pour l'analyse.

Malgré un intérêt croissant ces dernières années pour la recherche sur les interactions multimodales, notamment dans le domaine des EIAH (Dyke et al., 2007 ; Harrer et al., 2007), il existe encore peu de logiciels pour l'alignement et la représentation des données recueillies, la plupart proposant des formats propriétaires de structuration des données secondaires, ce qui rend difficile le partage des données entre les chercheurs. En outre, les phénomènes analysés étant complexes, les chercheurs ont souvent besoin de pouvoir coupler leurs premières analyses à des analyses d'un autre niveau ou de pouvoir comparer les analyses entre elles.

Le logiciel Tatiana, utilisé dans le projet Mulce pour l'analyse des interactions multimodales en ligne, a pour objectif d'aider les chercheurs dans l'analyse des corpus de traces d'interactions en ligne. Il permet de rejouer un ensemble d'éléments tracés qui ont été recueillis et synchronisés afin de comprendre, d'annoter et de coder tout ou partie de séances choisies en fonction de critères définis par les chercheurs. L'intérêt de l'utilisation de ce logiciel pour la recherche sur la multimodalité est qu'il s'inscrit dans une démarche *itérative* de l'analyse, permettant de créer autant d'artefacts d'analyse que nécessaires pour répondre aux questions de recherche (Lund et Milles, 2009), sous forme de représentation graphique (visualisation) et de catégorisation de la transcription. Les analyses effectuées peuvent ensuite être partagées et exportées (Dyke et al., 2008).

A partir du corpus global de Copéas, nous avons isolé deux phénomènes récurrents dans les interactions multimodales en ligne : les phénomènes de polyfocalisation (Jones, 2004) et, partant, de négociation de contexte, et les phénomènes de production langagière mettant en jeu plusieurs modalités (ici autour de l'écriture multimodale, en comparant deux sous-groupes de même niveau de L2). Nous avons ainsi constitué trois corpus distinguables "thématiques", relevant d'événements de communication caractéristiques de la communication multimodale en ligne (Ciekanski et Chanier, 2009a, 2009b, 2009c).

Le premier corpus distinguable porte sur un extrait du travail d'un sous-groupe de 3 apprenants du groupe des faux-débutants. Lors de cet épisode, ces derniers évaluent les premiers éléments de réponse apportés collectivement à la première question d'un quizz sur la notion d'interactivité dans le traitement de texte de l'environnement audio-graphique synchrone. Au bout de quelques minutes, le tuteur revient dans la salle où travaille le sous-groupe et intervient dans le clavardage pour apporter une aide linguistique aux apprenants. Ces interventions restent pourtant ignorées des apprenants qui poursuivent leur évaluation. Au bout de plusieurs tentatives sans résultat, le tuteur quitte la salle. La visualisation de cet épisode obtenue à partir de Tatiana permet de rendre visible la répartition des modalités entre participants (apprenants et tuteur). Il apparaît que les apprenants utilisent exclusivement les modalités audio et traitement de texte, alors que le tuteur intervient exclusivement dans le clavardage. Le clavardage étant pourtant visible par tous, les interactions qui ont lieu dans cet outil ne sont pas lues par les apprenants. Ainsi, loin de concevoir le contexte d'interaction comme donné par l'organisation spatiale de la plateforme, les apprenants semblent négocier leur espace de communication et de travail au fil de leurs échanges. Pour comprendre la façon dont se construit le contexte lors de cet épisode, nous avons catégorisé à l'aide de Tatiana la transcription multimodale en 5 catégories, rendant compte du processus de contextualisation propre à l'épisode analysé. Les interventions du tuteur apparaissent alors catégorisées comme "hors contexte" (cf. figure 2). Dans la mesure où le tuteur va et vient entre différentes salles de la plateforme audio-graphique synchrone, il joue un rôle marginal dans la participation au contexte de travail en tant que tel, n'a pas négocié l'utilisation du clavardage avec les apprenants et n'a donc pas été inclus dans le contexte partagé.

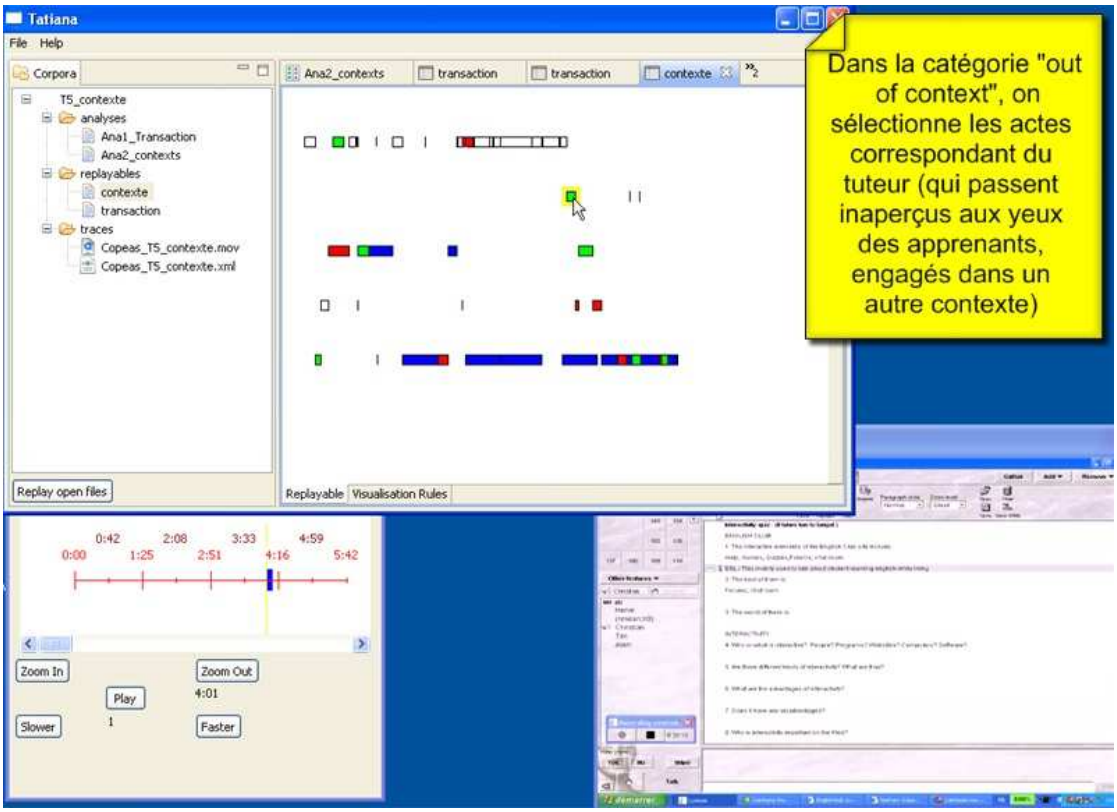

Figure 2- Interface de Tatiana montrant la synchronisation d'une visualisation chronologique des interactions filmées (en bas à droite de l'écran) avec le rejouable obtenu à partir des règles de visualisation du processus de catégorisation étudié dans l'extrait (en haut à gauche de l'écran).

Nous avons procédé de la même façon pour les deux corpus distinguables portant sur les phénomènes d'écriture multimodale, à partir du travail en parallèle dans deux salles différentes de deux sous-groupes (A et B) du groupe des faux-débutants, sur la même tâche rédactionnelle. La visualisation de chacun des deux épisodes révèle de grandes différences dans la manière dont les deux sous-groupes recourent à la multimodalité. La focalisation sur les acteurs met en exergue le fait que l'utilisation de la multimodalité dépend de stratégies individuelles qui ont à être négociées au sein du groupe pour rendre la collaboration efficace. L'utilisation de la multimodalité est également contextuelle dans la mesure où certains acteurs modifient leur "densité modale" (Norris, 2004) par rapport aux précédentes sessions. La compréhension de l'usage de la multimodalité pour produire un texte en L2 nécessite une analyse plus fine du processus rédactionnel. Pour ce faire, nous avons eu recours aux étapes du processus rédactionnel tel que décrit par (Zimmerman, 2000) : formulation, transcription, correction, reformulation et révision. Nous avons catégorisé dans un premier temps la transcription multimodale et recoupé ces informations avec l'utilisation que font les apprenants de la multimodalité lors de leur rédaction. Pour le sous-groupe A, nous notons que la coexistence de deux modalités pour l'écrit (le clavardage et le traitement de texte) permet de répondre à des fonctions distinctes, selon les apprenants. Ainsi, l'apprenant AT6 semble privilégier le clavardage au traitement de texte pour participer à la formulation et la correction des réponses, sans doute car ce dernier revêt un caractère moins permanent que le texte rédigé directement dans l'outil "traitement de texte". En effet, la production dans le clavardage ponctue les interactions qui se déroulent dans les autres modalités (audio et traitement de texte) et n'est pas au centre du discours. Cela permet une prise de risque moins

importante par rapport au fait d'écrire directement dans le traitement de texte. Pour le sous-groupe B, on notera qu'à la différence des apprenants du sous-groupe A les apprenants n'utilisent pas le clavardage. Le processus rédactionnel se réalise uniquement dans les modalités audio et texte. En outre, dans la mesure où le sous-groupe B fait un usage inhabituel de la L1 lors de cet épisode, nous avons catégorisé, dans un deuxième temps, l'usage de la langue (L1 et L2) à partir de la première catégorisation sur le processus rédactionnel décrit ci-avant. Cet étayage en L1 concerne principalement l'apprenant AT5, dans les phases de formulation et de révision, cette formulation en L1 étant ensuite prise en charge par un autre apprenant qui reformule et transcrit en L2 dans le traitement de texte.

Le travail sur les interactions multimodales en ligne montre qu'une compréhension de ses phénomènes ne peut se faire que dans un va-et-vient entre plusieurs niveaux de description (*intra-corpus*) et plusieurs corpus (*inter-corpus*) (section 3.2). Les différents niveaux de description requièrent une palette d'outils variés, en fonction des phénomènes étudiés. En effet, l'analyse de processus de groupe demande par exemple de « mesurer » les chevauchements des actes, de calibrer si des échanges représentent des interactions plus ou moins fortes, etc. Ces questions de recherche nécessitent de disposer d'outils traitant des données temporelles ainsi que des outils pour la reconnaissance de forme : reconnaissance de motifs d'actions (de patrons), reconnaissance de contenus d'échanges parlant d'un même « thème », comme le permettraient des outils de traitement automatique du langage (TAL). Par ailleurs, la compréhension des phénomènes étudiés ne peut être fine que si elle confronte plusieurs corpus de même type (en faisant varier les niveaux, les tâches, les modalités de collaboration, etc.) et de types différents (en faisant varier les environnements par exemple). C'est au prix d'une réflexion sur les formalismes et sur l'intérêt de tous au partage des données de recherche que l'étude des interactions en ligne en situation d'apprentissage permettra de réelles avancées sur le plan théorique et, partant, sur le plan des dispositifs d'apprentissage.

5 Conclusion

La notion de corpus d'apprentissage (LETEC) a été définie au regard des habitudes de recherche dans le domaine EPAL et des développements récents dans les différentes communautés s'intéressant à la recherche sur corpus. Le paradigme corpus comporte les quatre points indissociables suivants : le recueil systématique des documents liés à l'objet d'étude, la description du contexte, l'organisation et l'instrumentalisation en vue de traitements et les dispositions à prendre en vue de l'échange et du partage. Un corpus d'apprentissage (LETEC) assemble donc de façon systématique et structurée un ensemble de données, particulièrement d'interactions, et de traces issues d'une expérimentation de formation partiellement ou totalement en ligne, enrichies par des informations techniques, humaines, pédagogiques et scientifiques permettant leur analyse en contexte.

Dans le projet Mulce, les principes fondamentaux de la composition et de la structuration d'un corpus ont été mis en œuvre. La dimension éthique et juridique y joue un rôle essentiel puisqu'elle doit protéger les acteurs et contraintes les usages, conditions sans lesquelles les données ne seraient pas partageables. Ainsi les deux corpus de formations présentées dans l'article ont été systématiquement anonymisés. Depuis 2008, la structure opérationnelle Mulce-Struct a été affinée pour rendre possible diverses fouilles, recherches, étiquetages sur les interactions verbales (textuelles ou audio), ainsi que l'alignement entre les données audio ou vidéo et les transcriptions disponibles. Nous avons ainsi montré à partir de corpus distinguables issus de Simuligne et de Copéas l'intérêt des efforts de structuration permettant d'offrir au chercheur, accédant au serveur de la banque de corpus Mulce, un environnement

de travail lui permettant d'effectuer des fouilles intra ou inter corpus. Le second intérêt est de pouvoir réutiliser aisément ces données structurées pour mener des analyses avec des outils développés par la communauté de recherche sur les interactions en ligne. Le chercheur disposera donc dans Mulce d'un ensemble comportant une description structurée du corpus, contextualisé par rapport au corpus global (sous forme de commentaires libres et d'index précis renvoyant sur chacune des sous-parties d'un corpus global), des outils d'analyse associés et un ensemble de données prêtes à l'analyse ou contenant déjà des résultats d'analyse. Enfin, des liens relient un corpus distinguable à son corpus global et, le cas échant, à d'autres corpus distinguables pour des analyses inter-corpus. Actuellement, Mulce propose une trentaine de corpus distinguables prêts à l'analyse, relevant des trois types de corpus distinguables illustrés dans l'article.

Les efforts de standardisation convainquent de plus en plus de chercheurs dans nos communautés (EPAL09, CSCL09) et offrent donc une voie possible à ceux qui désirent partager leurs données et confronter leurs analyses. Notre proposition de structuration permettant d'inclure et de documenter un corpus avec toutes ses spécificités offre de travailler à des niveaux différents d'analyse, à partir d'un ensemble élargi d'interactions en ligne, comprenant des interactions synchrones et asynchrones, ce qui est encore peu souvent le cas. Reste la question du coût d'organisation et de structuration de tels corpus pour devenir partageables. Ils ne peuvent le devenir que dans des projets de recherche incluant plus d'une équipe de recherche dans une perspective interdisciplinaire des interactions en ligne. La prochaine étape passera sans doute par des collaborations pour transformer des ensembles de données recueillis par d'autres chercheurs dans le format d'un corpus d'apprentissage, lors d'un travail conjoint entre le collecteur et le créateur de corpus. Cette étape permettra d'estimer le coût du travail nécessaire à la transformation. Toutefois, gageons qu'elle permettra également un gain de temps important, étant donné la possibilité d'analyses multiples ainsi offerte, difficilement envisageable autrement, et finalement peu coûteuse en termes de structuration (transformation de format). Les chercheurs devraient également être gagnants en termes de reconnaissance de ce travail. Certes, l'existence des corpus d'apprentissage sera bénéfique pour l'ensemble de la communauté travaillant sur les interactions en ligne, mais elle n'en sera pas moins bénéfique au plan individuel dans une démarche d'évaluation des travaux scientifiques qui inclut celle des publications selon les procédures habituelles et des données associées.

Le projet Mulce s'achèvera dans sa première phase en 2010. De nombreuses pistes restent ouvertes en ce qui concerne les formalismes et les outils pertinents pour l'analyse des interactions en ligne, en fonction de leur nature (conversion au format TEI pour les données multimodales) et des outils de communication utilisées (les blogues par exemple). Ces pistes seront réellement pertinentes lorsque la communauté de chercheurs investira la plateforme Mulce en déposant ses données et en confrontant ses analyses avec celles d'autres chercheurs. Tel sera l'enjeu pour Mulce dans les mois à venir.

Bibliographie

Tous les liens Internet de cette section ont été vérifiés en date du 30 juin 2009.

AUDRAS I. et CHANIER T., 2008, « Observation de la construction d'une compétence interculturelle dans des groupes exolingues en ligne », dans *Apprentissage des Langues et Système d'Information et de Communication (Alsic)*, 11(1). pp. 175-204 http://alsic.revues.org/index865.html

BALDRY A. et THIBAULT P., 2006, *Multimodal Transcription and Text Analysis, a multimedia toolkit and coursebook with associated on-line course*, Equinox, Londres.

BASHARINA O. K., 2007, « An Activity Theory Perspective On Student-Reported Contradictions In International Telecollaboration», dans *Language Learning & Technology*, 11 (2), pp. 104-127, http://llt.msu.edu/vol11num2/basharina/

BAUDE O., BLANCHE-BENVENISTE B., CALAS M.F., CORDEREIX P., DE LAMBERTERIE I., GOURIE L., JACOBSON M., MARCHELLO-NIZIA C. et MONDADA L. (dir.), 2005, *Guide des bonnes pratiques pour la constitution, exploitation, conservation et diffusion des corpus oraux*, Editions du CNRS et DGLF-LF, Paris, http://www.culture.gouv.fr/culture/dglf/Guide_Corpus_Oraux_2005.pdf

BELZ J.A. et VYATKINA N., 2008, « The Pedagogical Mediation Of A Developmental Learner Corpus For Classroom-Based Language Instruction», dans *Language Learning & Technology*, 12(3), pp. 33-52, http://llt.msu.edu/vol12num3/belzvyatkina/

BETBEDER M.-L., CIEKANSKI M., GREFFIER F., REFFAY C. & CHANIER T., 2008, « Interactions multimodales synchrones issues de formations en ligne : problématiques, méthodologie et analyses», dans *numéro spécial EPAL (échanger pour apprendre en ligne), Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF)*, 15, J. Basque et C. Reffay (dir.), http://sticef.univ-lemans.fr/num/vol2008/06-betbeder/sticef_2008_betbeder_06.htm

BRUILLARD E., 2008, "Teacher development, discussion lists and forums : issues and results". In Mcferrin, K., Weber, R., Carlsen, R., & Willis, D.A. (eds). *Proceedings of Society for Information Technology and Teacher Education International Conference, SITE 2008*. Chesapeake, USA : AACE. p. 2950-2955.

CALICO, 2009, *Site où sont déposés des forums de discussion, avec les outils associés pour leurs analyses* [site Internet]. ERTé Calico. http://wims.crashdump.net/www/calico/

CATCOD, 2008, *Site de la communauté "Catalogage et codage de corpus oraux"* [site Internet]. Réseau Risc/CNRS. http://www.catcod.org/

CHANIER T., 2009, *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse des forums de Simuligne* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/mce-simu-forum.xml

CHANIER, T., 2004, *Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche ?* L'Harmattan. http://archivesic.ccsd.cnrs.fr/sic_00001103/fr/

CHANIER, T., CARTIER, J, 2006, "Communauté d'apprentissage et communauté de pratique en ligne : le processus réflexif dans la formation des formateurs", *Revue internationale des technologies en pédagogie universitaire (RITPU)*, 3(3), pp. 64-82. http://www.profetic.org/revue/IMG/pdf/RITPU-Vol_3_3.pdf

CHANIER, T., VETTER. A., 2006, "Multimodalité et expression en langue étrangère dans une plate-forme audio-synchrone". *Apprentissage des langues et Système d'Information et de Communication (Alsic)*, vol. 9. pp 61-101. http://alsic.revues.org/index270.html

CIEKANSKI, M., CHANIER, T., 2009a, *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse pour la notion de contexte dans Copéas* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-T5_contexte.xml

CIEKANSKI, M., CHANIER, T., 2009b, *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse pour la notion d'écriture multimodale dans Copéas* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-T8_ecriture_multimodale_s101.xml

CIEKANSKI, M., CHANIER, T., 2009c, *Manifeste du corpus distinguable fournissant des données prêtes à l'analyse pour la notion d'écriture multimodale dans Copéas* [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-T8_ecriture_multimodale_s102.xml

CIEKANSKI, M., CHANIER, T, 2008, "Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment". *Recall*, 20(2), pp. 162-182. doi:10.1017/S0958344008000426 http://edutice.archives-ouvertes.fr/edutice-00200851

CLAPI, 2009, Site *de la banque de corpus sur les interactions verbales* [site Internet]. Université Lyon 2 / Cnrs. http://clapi.univ-lyon2.fr

CRAPEL, 2007, Site du colloque *"Des documents authentiques oraux aux corpus : questions d'apprentissage en didactique des langues"*, décembre, Nancy [site Internet]. Nancy : ATILF. http://www.atilf.fr/atilf/evenement/Colloques/Crapel2007/crapel2007.htm

DATAVERSE, 2009, *Site de la communauté Dataverse Network développant un réseaux de banques de données de recherche associées aux publications*. [site Internet]. Harvard University. http://thedata.org/

DETEY, S., TCHOBANOV, A., DURAND, J., LAKS, B & LYCHE, C., 2007, "Des corpus oraux authentiques à leur exploitation didactique : accessibilité et prédidactisation des données pour l'enseignement du français parlé contemporain dans l'espace francophone : le projet PFC-EF". *Colloque (Crapel, 2007).* Résumé en ligne à http://www.atilf.fr/atilf/evenement/Colloques/Crapel2007/Resume_DETEY-TCHOBANOV-DURAND-LAKS-LYCHE.pdf

DYKE, G., GIRARDOT, J-J., LUND, K., CORBEL, A., 2007, "Analysing face-to-face computer-mediated interactions". *EARLI'07*. Budapest, Hongrie.

DYKE, G. LUND, K., GIRARDOT, J-J., 2008, "Managing, synchronising, visualising, analysing and sharing multimodal computer-mediated human interaction data: introducing Tatiana (A Trace Analysis Tool for Interaction Analysts)", ICLS *2008 Workshop: A Common Framework for CSCL Interaction Analysis*, Utrecht, Pays Bas, 23-28 Juin.

GARY KING, 2007, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," Sociological Methods and Research, 32(2), pp. 173--199, http://gking.harvard.edu/files/dvn.pdf

GOODWIN, C., DURANTI, A., 1992, "Rethinking context: an introduction". In Duranti, A et Goowin, C. (dir.) *Rethinking context, Language as an interactive phenomenon*, Camridge University Press, Cambridge, pp. 1-42.

HALLIDAY, M.A.K., 1989, "Part A". In Halliday, M.A.K. & Hasan, R. (dirs.). *Language, Context, and Text : Aspects of language in a social-semiotic perspective.* Oxford University Press. pp. 55-79.

HARRER, A., ZEINI, S., KAHRIMANIS, G., AVOURIS, N., MARCOS, J-A., MARTINEZ-MONES, A., MEIER, A., RUMMEL, N., SPADA, H., 2007, "Towards a flexible model for computer-based analysis and visualisation of collaborative learning activities". *Proceedings CSCL 2007*, 16-27 july 2007, New Jersey. USA.

IMS, 2008, *Site du IMS Global Learning Consortium, Inc* [site Internet]. http://www.imsglobal.org/ .

JEPSON, K., (005, "Conversations-and negotiated interaction- in text and voice chat rooms". *Language Learning and Technology*, 9 (3), pp. 79-98, http://llt.msu.edu/vol9num3/pdf/jepson.pdf

JONES, R., 2004, "The problem of context in computer-mediated communication". In LEVINE, P., SCOLLON, R. (dirs). *Discourse and technology: multimodal discourse analysis*. Georgetown University Press. pp. 20-33.

KERN, R., 2000, *Literacy and Language Teaching*. Oxford University Press.

KERN, R., WARE, P., WARSHAUER, M., 2004, "Crossing frontiers: new directions in online pedagogy and research". *Annual Review of Applied Linguistics*, vol. 24, pp. 243-260.

KRAMSCH, C., THORNE, S. L., 2001, "Foreign language learning as global communicative practice" . In D. Block & D. Cameron (dir.), *Globalization and language teaching*, pp. 83-100. Routledge : Londres.

LAMY, M-N., 2007, "Multimodality in online language learning environments: looking for a methodology". In Baldry, A., Montagna, E. (dirs). *Interdisciplinary perspectives on multimodality : theory and practice*. Proceedings of the third international conference on multimodality. Campobasso: Palladino. pp. 237-254.

LAKS, B., 2008, "Pour une phonologie de corpus". In *Le français à la lumière des corpus*, Durand, J. (dir.), numéro thématique de *Journal of French Language Studies*, 18 (1), pp. 3-32.

LEWIS, T., 2009, *Manifeste du corpus distinguable fournissant les données liées à l'article* (Lewis, 2006) [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/copeas-reflexive-tutor.xml

LEWIS, T., 2006, "When Teaching is Learning: A Personal Account of Learning to Teach Online". *Calico*, 23(3), pp 581-600, http://calico.org/html/article_110.pdf

LUND, K., MILLE, A., 2009, "Traces, traces d'interactions, traces d'apprentissages : définitions, modèles informatiques, structurations, traitements et usages". In Lund, K., Mille, A. (dirs). *Analyse de traces et personnalisation des environnements informatiques pour l'apprentissage humain*. Hermès. pp. 21-66.

MICASE, 2009, *Site de la banque de corpus "Michigan Corpus of Academic Spoken English"* [site Internet]. The University of Michigan. http://quod.lib.umich.edu/m/micase/

MULCE, 2009, *Site du projet Multimodal Learning Corpus Exchange* (2007-2010) [site Internet]. http://mulce.org

NORRIS, S., 2004, "Multimodal discourse analysis: a conceptual framework". In Levine, P., Scollon, R. (dirs). *Discourse and technology: multimodal discourse analysis*. Georgetown University Press. pp. 101-115.

OATES, J., 2006, "Ethical frameworks for research with human participants". In Stephen Potter (dir) *Doing Postgraduate Research*. Londres : Sage. pp. 200-228.

O'KEEFE, A. MCCARTHY, M. & CARTER, R., 2007, *From corpus to classroom. Language use and language teaching*. Cambridge University Press.

OLAC, 2008, "Best Practice Recommendations for Language Resource Description". In *Site de l'Open Language Archives Community*. University of Pennsylvania. http://www.language-archives.org/REC/bpr.html

PARPETTE, C., 2007, "Les discours universitaires oraux : questions de recherche, questions d'enseignement en FLE". *Colloque (Crapel, 2007).* Résumé en ligne à http://www.atilf.fr/atilf/evenement/Colloques/Crapel2007/Resume_PARPETTE.pdf

PÉREZ-LLANTADA, C, 2009, "Textual, Genre and Social Features of Spoken Grammar: A Corpus-Based Approach", *Language Learning & Technology*, 13 (1), pp. 40-58. http://llt.msu.edu/vol13num1/perezllantada.pdf

RASTIER F., 2005, "Enjeux épistémologiques de la linguistique de corpus". In Williams, G. (dir.) *La linguistique de corpus*. Presses Universitaires de Grenoble, pp. 31-46.

REFFAY, 2009, *Manifeste du corpus distinguable fournissant les données liées à l'article* (Reffay & Chanier, 2003) [corpus]. Mulce.org. http://mulce.univ-fcomte.fr/metadata/doc/visu-corpus/mce-simu-sna.xml

REFFAY, C., BETBEDER, M-L. (à paraître). "Sharing corpora and tools to improve interaction analysis. *EC-TEL 2009, Fourth European Conference on Technology Enhanced Learning*, Nice, septembre-octobre.

REFFAY, C., CHANIER, T., NORAS, M. & BETBEDER, M.-L., 2008, "Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche". In Basque, J. & Reffay, C. (dir.), *numéro spécial EPAL (échanger pour apprendre en ligne), Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (Sticef)*, vol. 15, http://sticef.univ-lemans.fr/num/vol2008/01-reffay/sticef_2008_reffay_01p.pdf

REFFAY C., CHANIER, T., 2003, "How social network analysis can help to measure cohesion in collaborative distance-learning". In *Procs. of Computer Supported Collaborative Learning Conference (CSCL'2003)*, Bergen, Norway, pp. 343-352, June, Kluwer Academic Publishers : Dordrecht (nl). http://edutice.archives-ouvertes.fr/edutice-00000422

SACODEYL, 2008, *Chaîne de traitements développée par le projet Sacodeyl*, corpus oraux d'adolescents européens composé à des fins pédagogiques [site Internet]. Espagne : Universidad de Murcia http://www.um.es/sacodeyl/en/pages/software.htm

SETTOUTI, L-S., PRIÉ, Y., MILLE, A., MARTY, J-C.,2006, "Systèmes à base de traces pour l'apprentissage humain". *Actes de TICE 2006*. Toulouse, octobre.

TATIANA, 2008, *Trace Analysis Tool for Interaction ANAlysts*.[logiciel] http://lead.emse.fr

THORNE, S. L., 2003, "Artifacts and cultures-of-use in intercultural communication", *Language Learning & Technology*, vol. 7(2), pp.38-67. . http://llt.msu.edu/vol7num2/thorne/

VETTER, A., CHANIER, T., 2006, "Supporting oral production for professional purpose, in synchronous communication with heterogeneous learners". *ReCALL*, 18(1), pp 5-23. http://edutice.archives-ouvertes.fr/edutice-00080316 doi:10.1017/S0958344006000218

WANG, Y., 2004, "Internet-based desktop videoconferencing in supporting synchronous distance language learning". *Language Learning and Technology*, 8 (3), pp. 90-121, http://llt.msu.edu/vol8num3/pdf/wang.pdf

ZIMMERMAN, R., 2000, "L2 Writing: subprocesses, a model of formulating and empirical findings". *Learning and Instruction*, 10, pp. 73-99.

Remerciements

Mulce est un projet soutenu par l'ANR Corpus et Outils en SHS (ANR-06-CORP-006). Il rassemble des membres des laboratoires LRL (Université Blaise Pascal), LIFC (Université de Franche-Comté) et CREET (The Open University), coordonnées respectivement par Thierry Chanier, Christophe Reffay et Marie-Noëlle Lamy, auxquels s'ajoutent Marie-Laure Betbeder et Maud Ciekanski.

Nous remercions F. Tajariol pour sa participation à la transformation des données du format Mulce au format Tatiana.

Notice biographique

Thierry Chanier est professeur des universités. Ses domaines d'enseignement et de recherche portent sur l'apprentissage des langues et les systèmes d'information et de communication, sur l'ingénierie de formation. Il étudie plus particulièrement les systèmes de formation à distance et les interactions en ligne sur des sujets tels que l'interculturel, le processus réflexif dans la formation des enseignants, le dialogue dans les environnements multimodaux. Il s'intéresse également à la structuration et aux modalités d'échanges entre chercheurs des corpus d'apprentissage. Mél. : thierry.chanier@univ-bpclermont.fr

Maud Ciekanski est maître de conférences en sciences du langage et chercheure à l'Université de Franche-Comté. Ses travaux relèvent de la didactique des langues et de l'analyse des interactions en situation d'apprentissage, dans les dispositifs d'apprentissage autodirigé avec soutien et dans les dispositifs d'apprentissage à distance. Depuis 2006, elle s'intéresse aux environnements d'apprentissage multimodaux et aux pratiques qui en découlent. Mél. : maud.ciekanski@univ-fcomte.fr